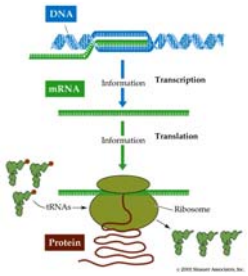
27622: Introduction to Bioinformatics, Turbo version

Henrik Nielsen, Associate professor

Center for Biological Sequence Analysis

What is bioinformatics?



What are bioinformaticians up to, actually?

- *Manage* molecular biological data
  - Store in *databases*, organise, formalise, describe...
- *Compare* molecular biological data
- Find *patterns* in molecular biological data
  - *phylogenies*
  - *correlations* (sequence / structure / expression / function / disease)

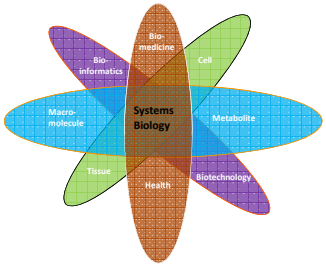
Goals:

- *characterise* biological patterns & processes
- *predict* biological properties
  - low level data ⇒ high level properties
    (eg., sequence ⇒ function)

Bioinformatics: neighbour disciplines

- Computational biology
  - *Broader concept: includes computational ecology, physiology, neurology etc...*
- -omics:
  - *Genomics*
  - *Transcriptomics*
  - *Proteomics*
- Systems biology
  - *Putting it all together...*
  - *Building models, identify control & regulation*

A view of Systems Biology

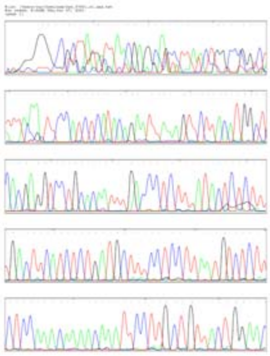


Bioinformatics: prerequisites

- **Bio- side:**
  - Molecular biology
  - Cell biology
  - Genetics
  - Evolutionary theory
- **-informatics** side:
  - Computer science
  - Statistics
  - Theoretical physics

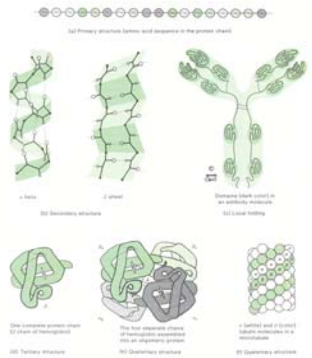
Molecular biology data...

- DNA sequences



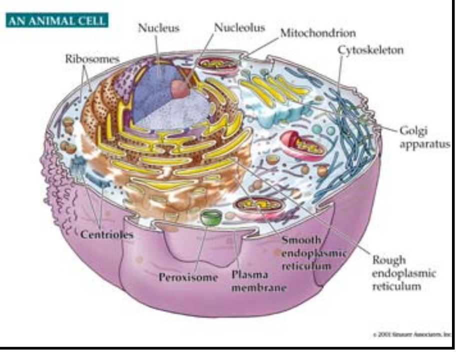
Molecular biology data...

- Amino acid sequences
- Protein structure:
  - X-ray crystallography
  - NMR



Cell biology & proteomics data...

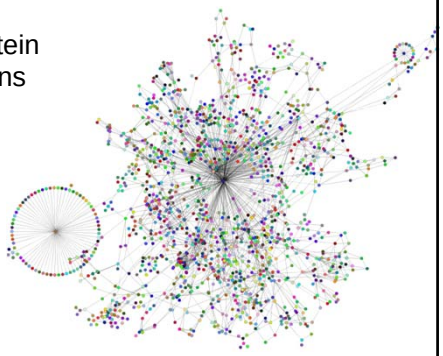
- Subcellular localization



CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

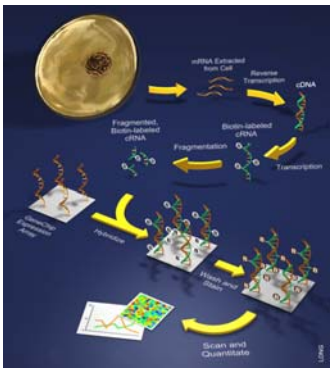
Cell biology & proteomics data...

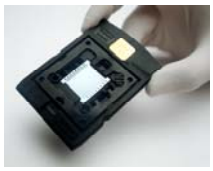
protein-protein interactions

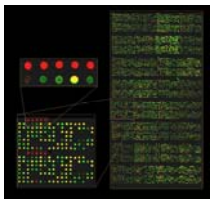


CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Transcriptomics: DNA microarray technology







CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Phenotype data: human diseases

Prediction methods

- Homology / Alignment
- Simple pattern (“word”) recognition
- Statistical methods
  - Weight matrices: calculate amino acid *probabilities*
  - *Other examples:* Regression, variance analysis, clustering
- Machine learning
  - Like statistical methods, but parameters are estimated by iterative *training* rather than direct calculation
  - *Examples:* Neural Networks (**NN**), Hidden Markov Models (**HMM**), Support Vector Machines (**SVM**)
- Combinations

The computer

- *Everything* can be reduced to bits (0 or 1)

Digital information

- A byte = 8 bits

0 1 0 0 0 0 0 1

Can be interpreted as

- The number 65
- The letter ”A”
- Part of a machine code instruction
- Part of a colour specification
- Part of a sound encoding
- …

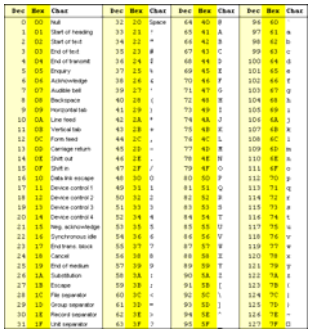
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Text files

A text file is a file where every byte is interpreted as a character

Examples

Plain text	.txt
Program settings	.ini
C source code	.c
Python script	.py
T_EX source	.tex
Web page source	.html
Sequences	.fasta

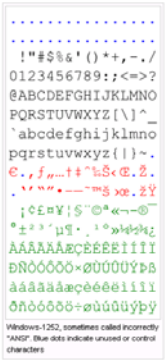


Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	⌂

The ASCII table

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Extended character sets



```
 !"#$%&'()*+,-./
0123456789:;<=>?
@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_
`abcdefghijklmno
pqrstuvwxyz{|}~.
```

Windows-1252, sometimes called incorrectly "ANSI". Blue dots indicate unused or control characters

The are *many* ways to interpret characters with values above 127. Here, you see two of them.

"Mac OS Roman" Encoding:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
032		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
048	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
064	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
080	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
096	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
112	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
128	Ä	Å	Ç	É	Ñ	Ö	Ü	á	à	â	ä	ã	å	ç	é	è
144	ê	ë	í	ì	î	ï	ñ	ó	ò	ô	ö	õ	ú	ù	û	ü
160	†	°	¢	£	§	•	¶	ß	®	©	™	´	¨	≠	Æ	Ø
176	∞	±	≤	≥	¥	µ	∂	∑	∏	π	∫	ª	º	Ω	æ	ø
192	¿	¡	¬	√	ƒ	≈	Δ	«	»	…		À	Ã	Õ	Œ	œ
208	–	—	“	”	‘	’	÷	◊	ÿ	Ÿ	⁄	€	‹	›	fi	fl
224	‡	·	‚	„	‰	Â	Ê	Á	Ë	È	Í	Î	Ï	Ì	Ó	Ô
240		Ò	Ú	Û	Ù	ı	ˆ	˜	¯	˘	˙	˚	¸	˝	˛	ˇ

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Text files—line endings

- UNIX standard (including Mac OS X):
  - 10 — LF ("Line feed" char).
- Old Mac (System 9 and before):
  - 13 — CR ("Carriage Return" char).
- DOS/Windows:
  - 13, 10 — both CR and LF.

A good text editor can handle all three systems.

Notepad for Windows *cannot!*

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

www.jedit.org

jEdit